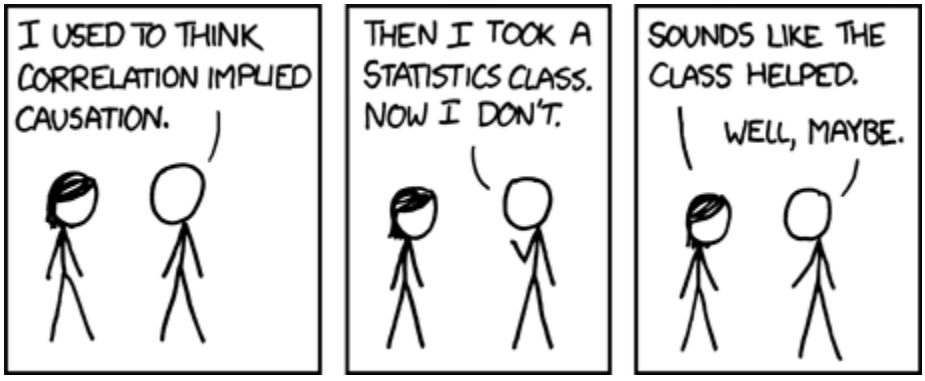
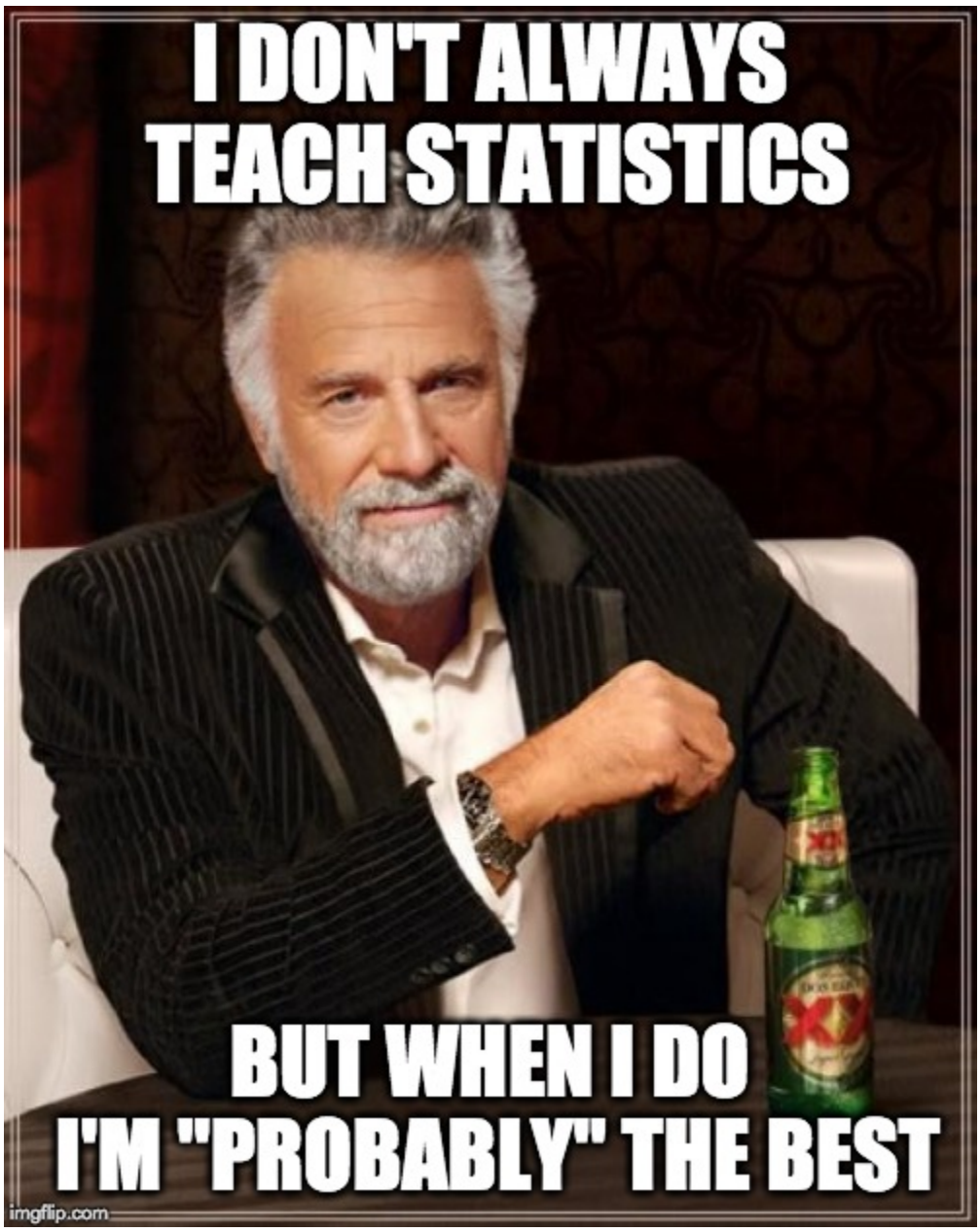
Unit 1 Live Session - W203, Statistics for Data Science (Summer'20)

1. What's statistics? What's this class?



2. Instructor Introduction



3. Our Students

We will ask you for short introductions on Slack. Question: What does a *data scientist* look like?

4. Weekly Workflow

A typical week of the course proceeds as follows:

1. **Before live session:**
    - **A. Watch realated async content.**
    - **B. Complete related required readings.**
2. **In live session:**
    - **A. We will build upon the async to test and extend your understanding**
3. **After live session:**
    - **A. complete the homework for the unit (due before next live session)**

5. Important Resources

- Course Webpage (https://w203-summer-20.github.io/class)
  - Up-to-date course policies
  - Calendar with weekly deliverables and deadlines
- ISVC (https://learn.datascience.berkeley.edu/login)
  - Platform with course videos
  - Submit your homework/labs here
    - One output file (either PDF or HTML)
    - One source file (IPYNB or RMD or R)
- Github Org: w203-summer-20 (https://github.com/w203-summer-20)
  - Source of homework, labs, and live session documents
  - Check an invite to our *github.org* to your iSchool email
    - it will allow you to join with any github account you like.
- Slack #w203_mids_sum_20 (https://ucbischool.slack.com/archives/C012FARHLJE)
  - Our forum for questions about content, announcements, etc.

6. Homework (HW)

- Almost weekly.
- Access HW1 via Github.Org (https://github.com/w203-summer-20).
  - Open the repository `unit\_1\_hw`.
  - If you are unfamiliar with *Github*, don't worry about the details right now. Find the green button to download a zip file containing all documents. *You can then open the notebook using Jupyter.
- Submit HW1 via `ISVC>Assessments`
  - There is a page there to upload your solution file

7. How to Succeed in this Class

Here are some helpful strategies:

1. **Get to know your readings, esp. *Devore* and *Wooldridge* textbooks**
    - **A. Some questions may come directly from readings**
2. **Do your best on assignments**
3. **Strategize about the HW with friends as much as you want, but submitted work must be your individual work.**
    - **A. Note: do not discuss quizzes/labs with anyone else unless otherwise specified**
4. **Try each problem on your own for at least a little bit.**
    - **A. We don't ask plug-and-chug problems in this class.**
    - **B. 99% of your time will be spent connecting words in English to mathematical objects.**
    - **C. It's normal not to know what formula to apply - once you do, you're almost done.**
    - **D. If lost:**
        - **a. review similar examples in textbook and async videos**
        - **b. write down relevant definitions**
        - **c. draw a relevant picture/diagram**
        - **d. express the unknowns in terms of known variables**
    - **E. If stuck on a problem for 30 min, talk to a classmate, come to OH, or make a post on class forum**
        - **a. When posting describe your approach, but do not give out solutions (whether correct or not)**
        - **b. Give other students a joy of solving the problem on their own**
5. **Form study groups!**
6. **Come to office hours (OH)! To all OH!**
    - **A. Instructors can prepare a specific example, if you send your in advance**

8. Software

You have 3 ways to use R code interpreter:

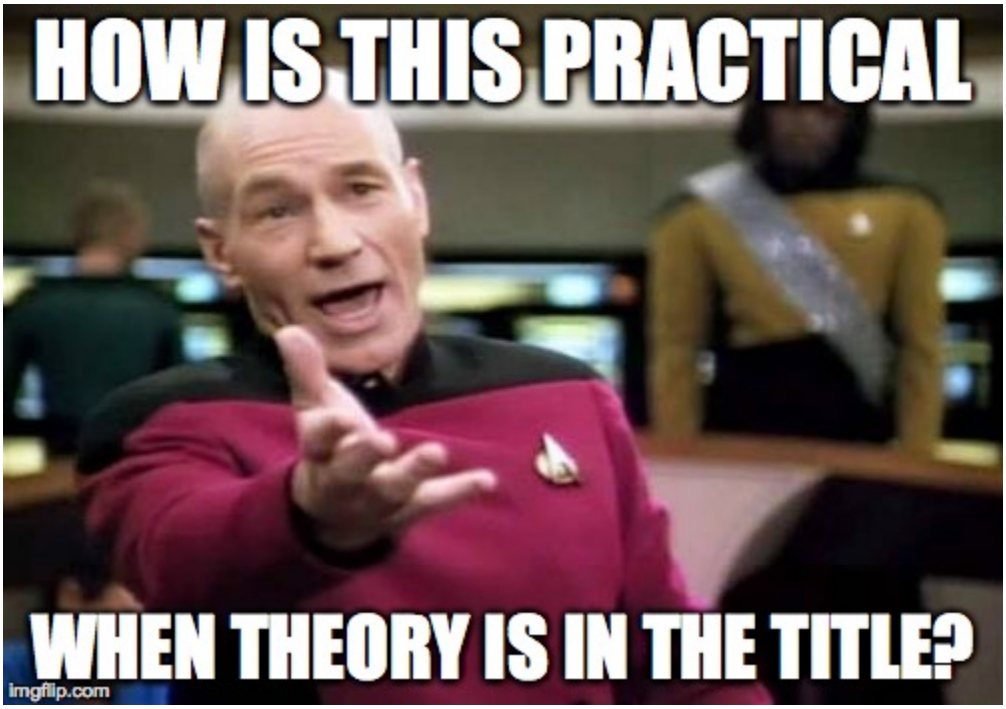
1. **R project (https://cloud.r-project.org) and RStudio (https://www.rstudio.com/products/rstudio/download)**
   - **A. Video tutorial (https://www.youtube.com/watch?v=9-RrkJQQYqY)**
2. **Anaconda, Jupyter Notebooks (https://www.youtube.com/watch?v=jZ952vChhuI) with an R kernel:**
   - **A. Video tutorial for Anaconda (https://www.youtube.com/watch?v=YJC6ldI3hWk)**
     - **a. It contains Jupyter Server and Notebook**
   - **B. Video tutorial for R Kernel (https://www.youtube.com/watch?v=SXBxKe8sK6I)**
     - **a. Just run a command `conda install -c r r-essentials` in CLI (https://en.wikipedia.org/wiki/Command-line_interface) of your operating system (OS)**
3. **Web-based RStudio using Berkeley's datahub (https://r.datahub.berkeley.edu)**
   - **A. Convenient, if you can't install software on your PC**

The first few weeks are heavy in math. Turn in as a single `.pdf` file! You can handwrite/scan or type up (preferred) your solutions with these tools:

1. **LyX (https://www.lyx.org/): Visual LaTeX. Excellent tool**
2. **Microsoft Word has a great formula editor**
3. **Google Doc has a reasonable formula editor**
4. **Any Latex editor and turn in the typesetted `.pdf`**

LaTeX is painful at first due to errors from failed compilations, but is useful in the long run. We will be able to use the syntax for communicating equations in the notes and chat pods.

9. Practical Probability Theory



Exercise: Sample Space

For each of the following experiments:

1. **Give example of an event**
2. **Define a state space Ω**
3. **How big is Ω (what is its cardinality, i.e. how many elements does it have?)**

Exercise: Rolling a 6 sided die

Consider a standard [52-card deck (https://en.wikipedia.org/wiki/Standard_52-card_deck)](https://en.wikipedia.org/wiki/Standard_52-card_deck) with 13 ranks and 4 suits.
Image source: [Wikipedia (https://en.wikipedia.org/wiki/Standard_52-card_deck)](https://en.wikipedia.org/wiki/Standard_52-card_deck)

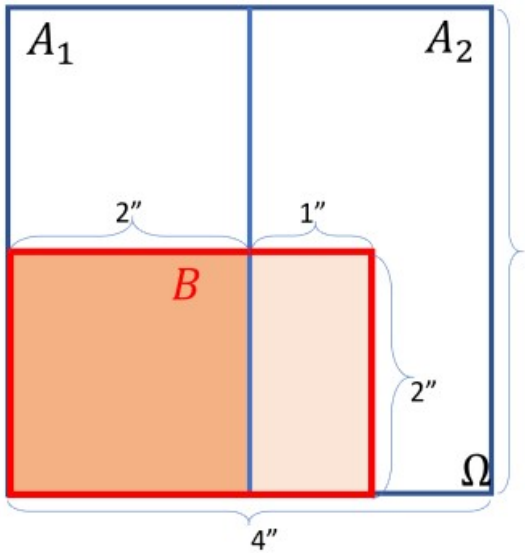
Example set of 52 playing cards; 13 of each suit clubs, diamonds, hearts, and spades

	Ace	2	3	4	5	6	7	8	9	10	Jack	Queen	King
Clubs													
Diamonds													
Hearts													
Spades													

Exercise: Drawing a 5 Card Poker Hand

Exercise: Choosing 1000 U.S. citizens for a survey

Exercise: Compute with LTP



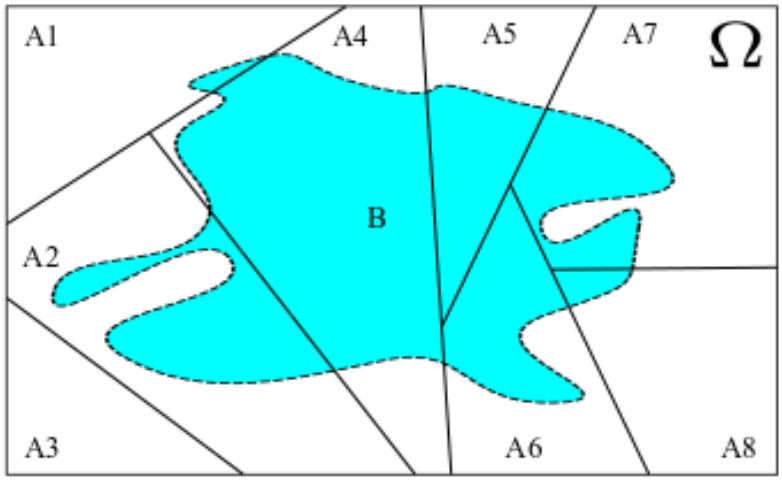
Law of Total Probability.
Consider sample space Ω as 2×2 square of points partitioned into halves by A_1, A_2. Also B is a subset of events as shown in figure. All events are equi-weighted.

The shapes are simple and we can directly compute the $\mathbb{P}B = \frac{2 \cdot 3}{4 \cdot 4} = \frac{3}{8}$, but sometimes it is easier to apply the Law of Total Probability:

$$\mathbb{P}B = \mathbb{P}[B|A_1] \cdot \mathbb{P}A_1 + \mathbb{P}[B|A_2] \cdot \mathbb{P}A_2$$

What are the probabilities in the summation?

Exercise: Explain LTP



Consider a partition of Ω state space 1. What would disjoint, but not mutually exclusive set of A_i's would look like? 1. What would mutually exclusive, but not disjoint set of A_i's would look like? 1. Explain what the Law of Total Probabilities (LTP) means intuitively.

$$P(B) = \sum_{i=1}^{N} P(A_i \cap B) = \sum_{i=1}^{N} P(B|A_i)P(A_i)$$

Exercise: Conditional Probabilities

Suppose you're taking a statistics class and that in each week you are either caught up or behind on the readings. It is difficult to incorporate a new class in to your schedule so the probability that you will be caught up in Week 1, having viewed all async and completed the pre class exercises before the live session, is 0.7. From then on, if you are caught up in a given week, the probability that you will be caught up in the next week is 0.7. If you are behind in a given week, the probability that you will be caught up in the next week is 0.4. What is the probability that you are caught up in week 3?
You can use a tree diagram or a formula (which one?) to solve the problem.

Exercise

A test for certain disease is assumed to be correct 95% of the time: if a person has the disease the test will give a positive result with probability 0.95. If a person does not have disease the test will give a negative result with probability 0.95. A random person drawn from a certain population has a probability 0.001 of having the disease. Given that a person drawn at random just tested positive, what is the probability that they have the disease? Setup:
Similar to above, we can either build a tree or state all probabilities symbolically. Let D be the event of a person having the disease, "$+$" is the event of the test showing positive, and "$-$" is the event of the test showing negative. Then we need to find $\mathbb{P}[D|+]$

10. Reminders

Before next week:

1. **Install Jupyter with an R kernel (or make sure you can use r.datahub.berkeley.edu)**
2. **Complete the Unit 1 homework located in our Github.org (due before next live session).**
3. **Watch all unit 2 async content**

In [ ]: